(12) UK Patent Application (19) GB (11) 2 339 656 (13) A

(43) Date of A Publication 02.02.2000

(21) Application No 9815300.0

(22) Date of Filing 14.07.1998

(51) INT CL[7]
G06F 1/00

(52) UK CL (Edition R )
H4P PDCSA
U1S S2209

(56) Documents Cited
EP 0798619 A US 5467447 A

(58) Field of Search
UK CL (Edition Q ) H4F FBB , H4P PDCSA PDCSK PDCSX
INT CL[6] G06F 1/00 , H04N 1/32
Online:WPI,INSPEC

(71) Applicant(s)
Laurence Frank Turner
4 Barrells Down Road, BISHOP'S STORTFORD, Herts, CM23 2SU, United Kingdom

Athanassios Manikas
139 Hounslow Road, HANWORTH, Middlesex, TW13 6PX, United Kingdom

(72) Inventor(s)
Laurence Frank Turner
Athanassios Manikas

(74) Agent and/or Address for Service
Reddie & Grose
16 Theobalds Road, LONDON, WC1X 8PL, United Kingdom

(54) Abstract Title
**Electronic text watermarking**

(57) An improved method of inserting code data into a text data file, or 'watermarking' is provided. Selected characters are overwritten or underwritten with one or more other characters in such a way that only the selected character is visible. Unlike other known methods, this method is insensitive to formatting changes, such as change of font or justification. The watermark can be encrypted.

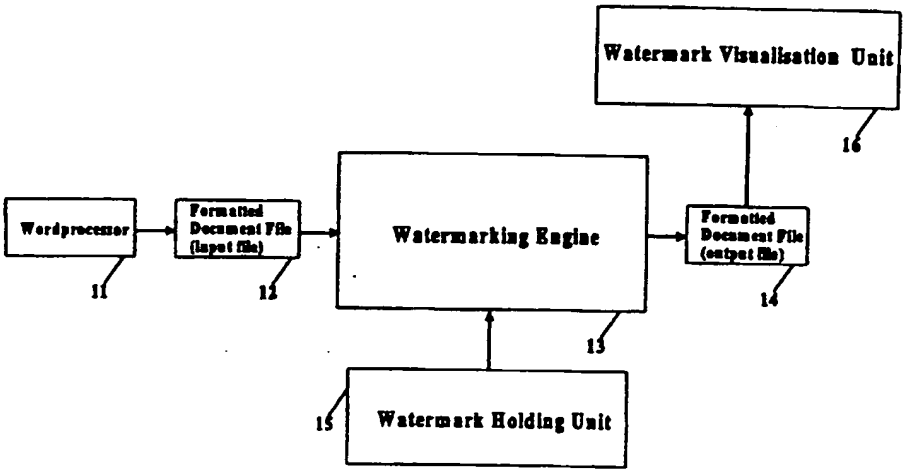


*Figure-1*

GB 2 339 656 A

At least one drawing originally filed was informal and the print reproduced here is taken from a later filed formal copy.

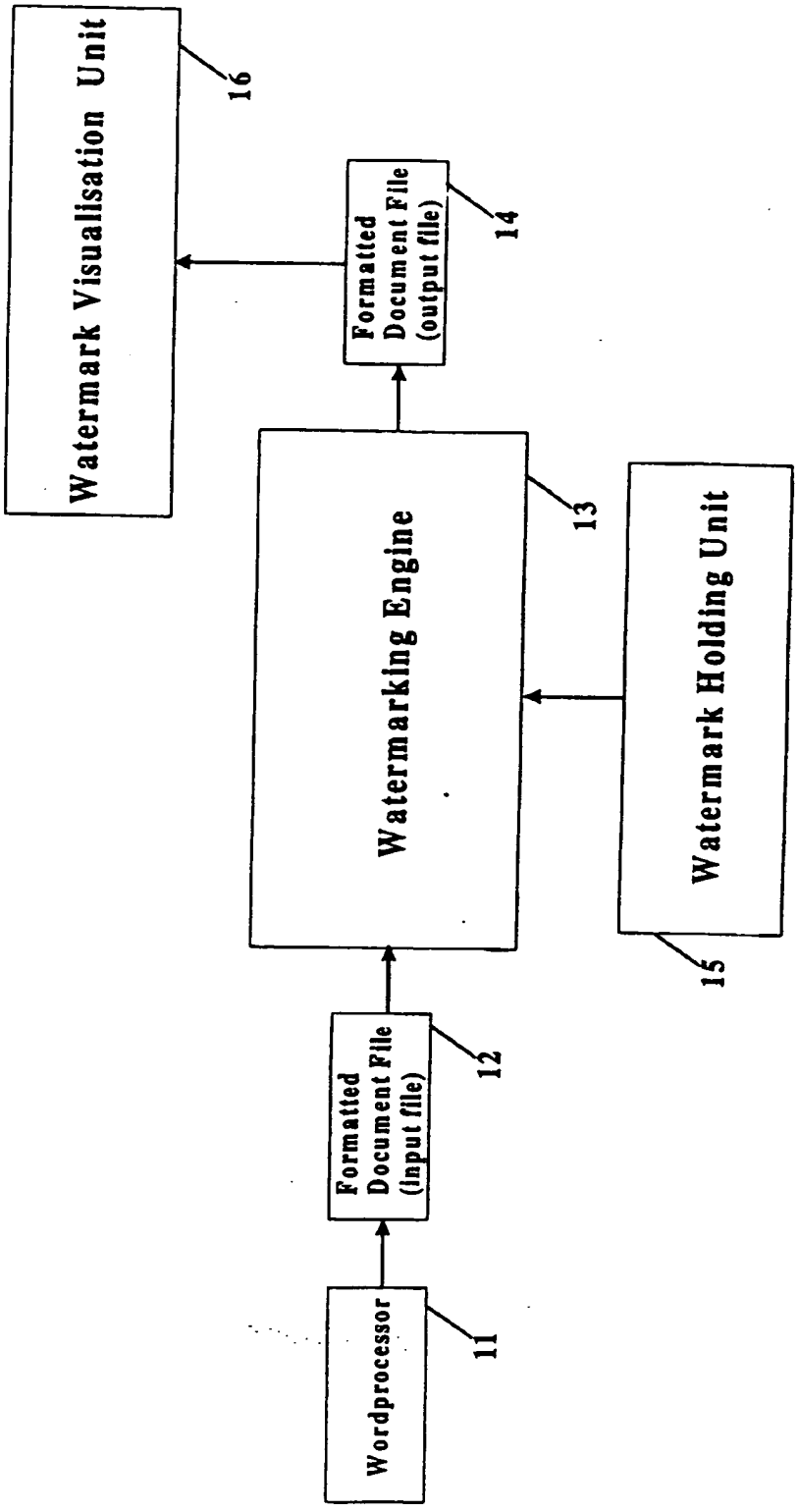
Watermark Visualisation Unit
16
Formatted Document File (output file)
14
Watermarking Engine
13
Watermark Holding Unit
15
Formatted Document File (input file)
12
Wordprocessor
11

*Figure-1*

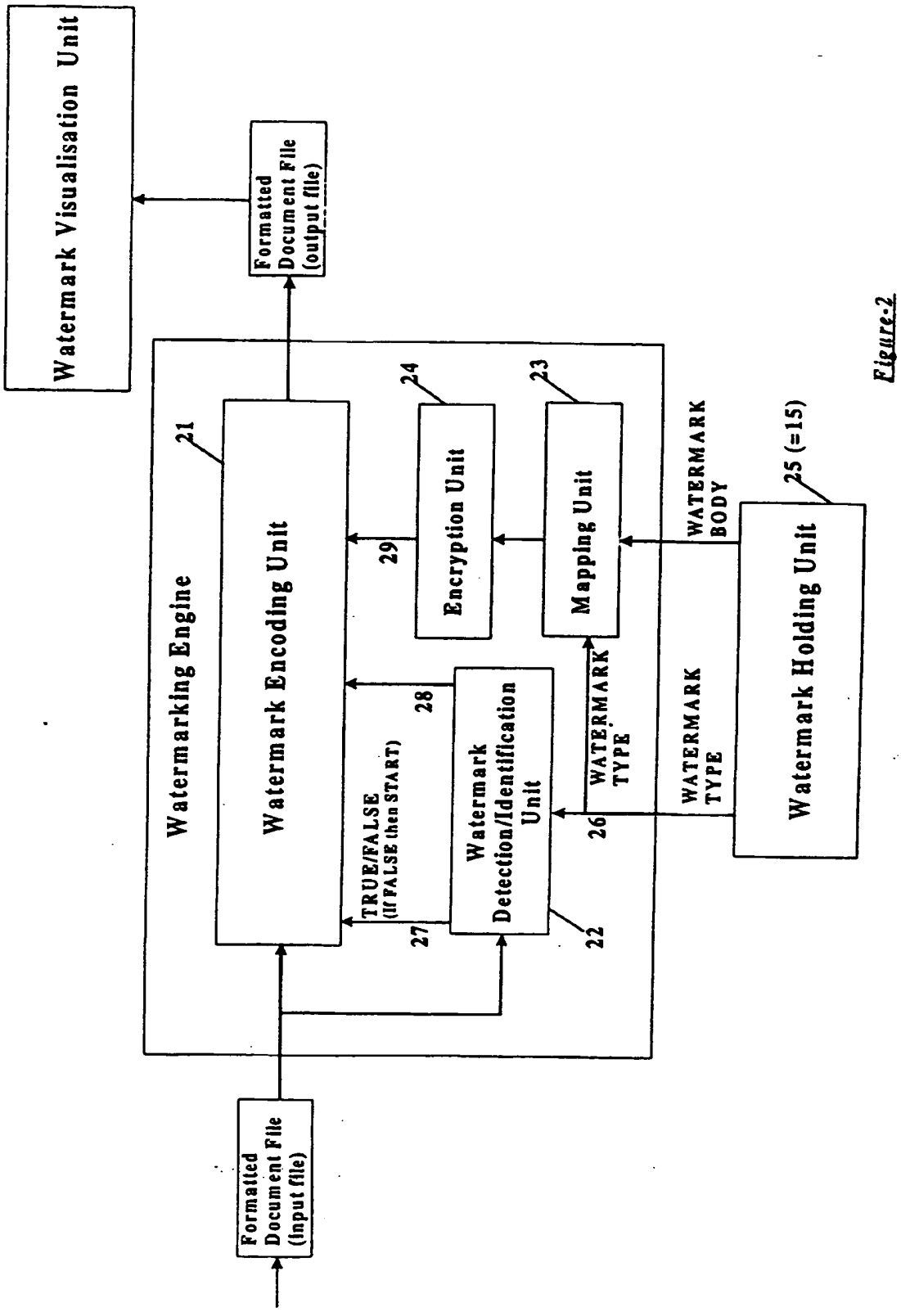
Watermark Visualisation Unit
Formatted Document File (output file)
Watermarking Engine
21
Watermark Encoding Unit
24
Encryption Unit
23
Mapping Unit
WATERMARK BODY
25 (=15)
Watermark Holding Unit
29
TRUE/FALSE (If FALSE then START)
28
Watermark Detection/Identification Unit
WATERMARK TYPE'
WATERMARK TYPE
27
22
26
Formatted Document File (input file)

*Figure-2*

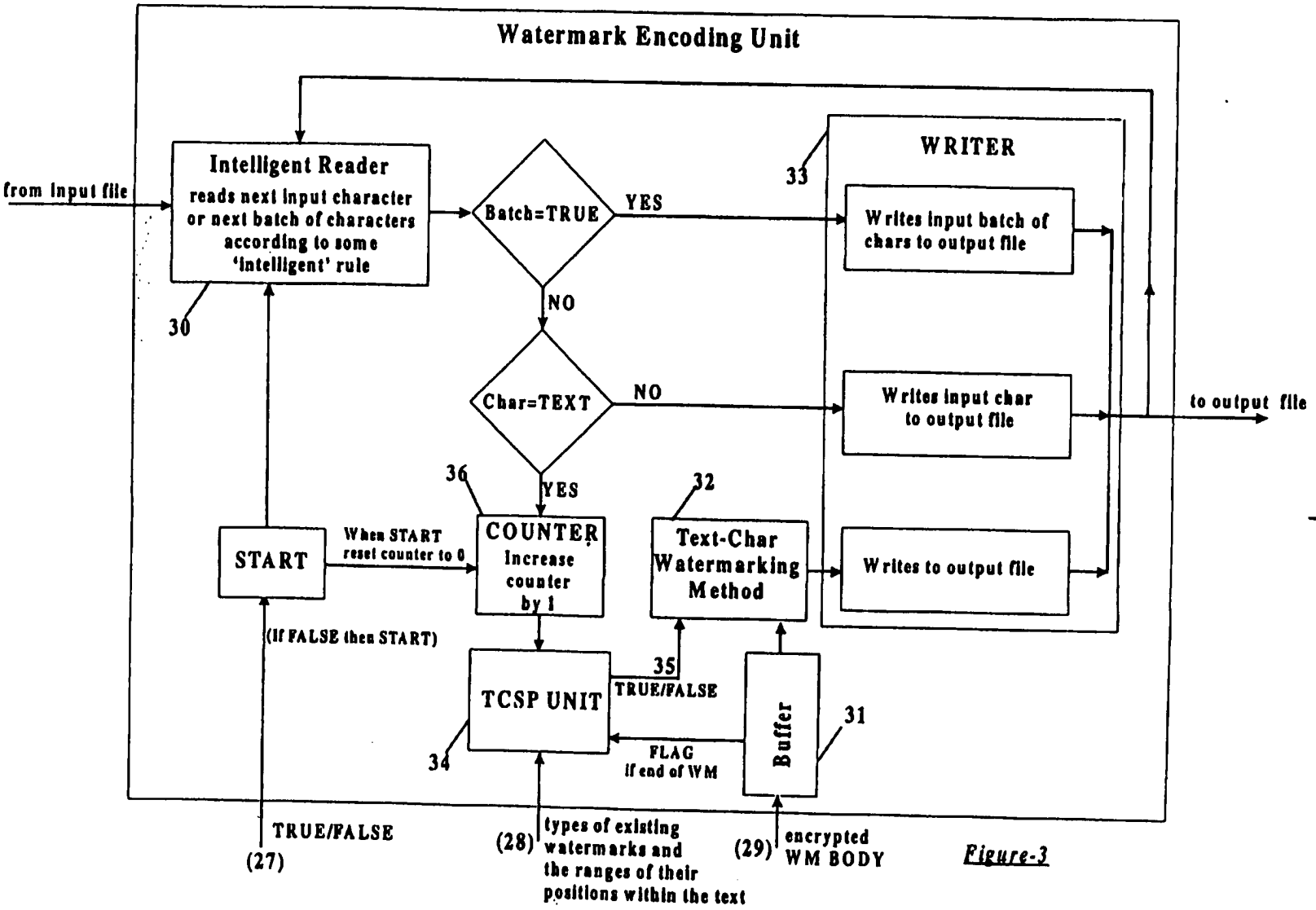
Watermark Encoding Unit
from input file
Intelligent Reader
reads next input character or next batch of characters according to some 'intelligent' rule
30
Batch=TRUE
YES
NO
WRITER
33
Writes input batch of chars to output file
Char=TEXT
NO
Writes input char to output file
to output file
YES
36
32
COUNTER
Increase counter by 1
START
When START reset counter to 0
(If FALSE then START)
Text-Char Watermarking Method
Writes to output file
TCSP UNIT
35
TRUE/FALSE
FLAG
If end of WM
Buffer
31
34
TRUE/FALSE
(27)
(28)
types of existing watermarks and the ranges of their positions within the text
(29)
encrypted WM BODY

*Figure-3*

# Electronic Watermarking

The invention relates to a method of inserting code data into text data, which is sometimes called watermarking a text.

With the increase in electronic trading and the delivery of text in electronic form there is a growing need for secure methods of identifying ownership of written texts and for identifying the parties involved in electronic transactions. This can be achieved using electronic watermarking. This involves the insertion of information into the text material. The information can later be detected, read and displayed and used to identify aspects such as ownership. The information is, ideally, inserted in such a way as to not affect the perceived quality of the text.

Watermarking of documents allows audit trails to be set up easily. For example, if a particular document had to be watermarked by each owner before that owner was allowed to sell on the document to someone else, then from the different watermarks in the document it would be possible to identify all the previous legitimate owners of the document.

Techniques for watermarking text material are known, for example see "Challenges for copyright in a digital age", I D Bramhill & M R C Sims, B T Technol J Vol 15 No 2 April 1997. These methods involve making small alterations to specific characters, for example slight broadening of a comma or full stop, producing physically different characters which appear unaltered to the naked eye. The differences can be picked up electronically, for example when the document is scanned. However, these techniques are extremely sensitive to even minimal formatting changes, such as changes of font and justification.

The invention provides an improved, more robust method for inserting code data into a text data file representing a string of characters. The invention is defined in the appended claims to which reference should now be made.

In the present invention a selected character is marked using overwriting or underwriting, that is, the placing of one or more other character(s) at the same position in the text as the selected character. These other characters might all be the same as the selected character, so that the under/overwritten characters cannot be seen. Alternatively, the other characters may include a 'space' or a 'null' character, or a combination of these overwritten one on top of another, making them invisible in the printed text. This combination of 'space' and/or 'null' characters is then overwritten with the selected text character. Thus the encoded material is inserted in such a way that it does not affect in any way the quality of the text.

Other characters which may be used in the set of characters with which each selected character is replaced may be colourless characters, including colourless underlining or overscoring characters, that is, characters which are at the same position in the string of characters as the selected character but which, in the text document, appear above or below the selected character.

The message or information to be inserted into the text is preferably encrypted and it is placed into the text in such a way as to make it difficult for unauthorised persons to remove the message, or watermark, from the text.

Preferred embodiments of the invention will now be described with reference to the drawings in which:

Figure 1 shows a schematic of a system according to one embodiment of the invention;

Figure 2 shows a Watermarking Engine in more detail; and

Figure 3 shows a block diagram of the structure of the Watermark Encoding Unit.

In the following description the term 'watermark' is used to mean a message which is to be inserted into the text. As is shown in Figure 1, the watermarking of a text-formatted file, which is referred to as the Input File 12 and which has been created using a Wordprocessor 11, is performed using a Watermarking Engine 13. The watermarked text-formatted file produced is referred to as the Output File 14. The Watermarking Engine 13 reads the Input File 12 and embeds, according to pre-specified rules, watermarks that are contained in a Watermark Holding Unit 15 and finally saves the result into the Output File 14. Inserted watermarks can be 'seen' using the Watermark Visualisation Unit 16.

Figure 2 shows, in general, the elements of a text watermarking system according to one embodiment of the invention. The Watermarking Engine 13, as shown in Figure 2, consists of a Watermark Encoding Unit 21, a Watermark Detection and Identification Unit 22, a Mapping Unit 23 and an Encryption Unit 24.

The Watermark Holding Unit 15, which is not part of the Watermarking Engine 13 itself, may contain a single watermark or a collection of a number of different watermarks. This unit 15 may be centralised or distributed throughout an electronic trading network. Centralised operation would enable some, or all, of the watermarks to be inserted by an authorised agent, while the distribution of the unit would make it possible for specified types of watermark to be inserted by an authorised party, or parties, over a network.

A watermark obtained from the Watermark Holding Unit 15 is defined by two parameters, the first parameter is the 'Type' of the watermark, and the second is the 'Body' of the watermark. The 'Type' is an identifier and the 'Body' is the content, comprising a sequence of digits, symbols etc. corresponding to, for example, an ISBN as used in publishing, or a name which may be the name of an author or distributor, or any such combination of alpha-numeric and special characters. Two illustrative examples of watermarks are given in Table 1.

Table 1

| | TYPE | BODY |
|---|---|---|
| Example 1 | 1st | 1-12345-333-5 |
| Example 2 | Client | Smith |

The sequence of characters representing the Type and Body of the watermark is taken from the Watermark Holding Unit 15 and fed to a Watermark Mapping Unit 23 where the body of the watermark is mapped into a longer sequence of characters. The mapping is dependant, not only on the Body, but also on the Type of the watermark. Through the mapping the Body of the watermark is converted, generally, into binary form. The invention is not restricted to the situation in which the output is a sequence of binary digits.

The output from the Mapping Unit 23 is input to the Encryption Unit 24 which operates in a manner well known to those in the art [see "Cryptography and Secure Communications" Rhee M Y, McGraw Hill Book Company, Singapore, 1994 which is incorporated herein by reference]. The output of the Encryption Unit 24 represents a binary, encrypted form of the watermark information that is to be embedded in the formatted text file. The Watermark Detection and Identification Unit, 22, examines the Input File 12 and, from Line 26, obtains information regarding the type of

watermark contained in the Watermark Holding Unit 25 and checks the Input File 12 for the presence of all watermark 'Types' and displays any that are found to exist in the Input File. If any of the found watermarks is of the same type as that which is currently to be embedded then a TRUE, hence 'inhibit', command is sent along Line 27 to the Watermark Encoding Unit 21 in order to prevent watermarking. Otherwise, a signal is sent to the Watermarking Encoding Unit 21 along Line 27 instructing the unit 21 to start its operation. Also, if rules exist for the order in which watermark types are to be inserted then any attempt to insert a watermark in violation of these rules is inhibited via Line 27.

The Watermark Encoding Unit 21 which is shown in general form in Figure 3, consists of five main sub-units: a Buffer 31 for holding the mapped, encrypted body of the watermark to be inserted, an Intelligent Reader 30 which reads the input file in a manner to be described later, a Text-Character Watermarking Unit 32, a Writer 33 which writes text characters, watermarked or otherwise, to the Output File and a Text-Character Selection Procedure Unit (TCSP Unit) 34.

The operation of the Watermark Encoding Unit 21, which is under the general control of the Text-Character Selection Procedure Unit 34 will now be described in general terms. The operation begins once a FALSE signal is received on Line 27 from the Watermark Detection and Identification Unit 22. Then the counter 36 shown in Figure 3 is reset to 0 and the Intelligent Reader 30 starts reading the Input File 12. The Intelligent Reader 30 reads the text file 12 according to its rules of operation which depend on the format and type of the input file and any other information available in respect to the input file. If the Intelligent Reader reads a batch of characters then they are written directly by the Writer 33 to the Output File 14. If the Reader 30 reads a

single character (a byte) then a check is made to determine whether this character is a text-character. If it is not a text character then it is again written directly to the output file. If, however, it is a text character then the Counter 36 is incremented by 1 and the TCSP Unit 34 takes over and, on the basis of a set of rules, controls whether or not this text character should be watermarked by the Text-Character Watermarking Unit. If the decision is that the text character should not be watermarked then the Text-Character Watermarking Unit 32 receives a FALSE command on Line 35 and the text character is written to the Output File. However, if the decision is that the text character should be watermarked then the command on Line 35 is TRUE and this instructs the Text-Character Watermarking Unit to perform the following two operations:

1) to read the next symbol of the mapped and encrypted Watermark Body from the Buffer 31;

2) watermark the selected text character in accordance with one of the watermarking methods to be described later.

The watermarked text character is written to the Output File and the control passes back to the Intelligent Reader 30 which then reads the next character, or a batch of characters.

It is clear from the above discussion that the Text-Character Selection Procedure Unit 34 controls the watermarking insertion and decides which characters should be watermarked according the TRUE or FALSE signal on Line 35 and a predefined set of rules, specific examples of which will be given later.

This decision taken by the TCSP Unit as to whether or not a text character should be watermarked is based on the following three general rules:

1. constraint rules;
2. short jumping rules;
3. long jumping rules.

The constraint rules are based on the information provided by the Detection and Identification Unit 22 to the TCSP Unit 34. This is comprehensive information as to the types of the existing watermarks within the text and the ranges of positions within which the watermarked text characters fall. The constraint rules become active every time the content of the counter is such that it falls within the constraint ranges referred to above. The command on Line 35 then becomes FALSE thereby prohibiting watermarking of any text character under consideration.

The short jumping rules, which come in to effect when the constraint rules are inactive, determine whether or not a text-character is to be watermarked by a digit of the specified watermark. In addition the short jumping rules determine the number of text characters that there is to be between successive watermarked characters. This separation may be by a predetermined number of text characters, or by an integer number generated at random, subject to a maximum value.

One such short jumping rule is that in which the counter content is evaluated modulo-M, where M is a predefined integer number, and the modulo-M value is then compared with some pre-selected integer number K which satisfies the condition $0 \leq K < M$. If the two numbers are the same, that is, (counter content modulo-M)=K, then the text character under consideration is selected to be watermarked and the Writer 33 is instructed to write the watermarked text character to the Output File 14. If the two integer numbers are not the same that is K≠(counter content modulo-M), then the text character is written to the Output File 14 without it being watermarked. In this case the watermarking digit under

consideration is kept ready for use until such time as another text-character is read and is selected to be watermarked. The process is repeated until all of the digits that make up the mapped and encrypted Body of the watermark have been inserted.

The long jumping rules are activated by the buffer when the last digit/symbol of the mapped and encrypted Watermark Body has been inserted into the text file. These rules are determined in a straight forward manner depending on the number of different types of watermarks to be inserted, the lengths of the watermarks, the frequency with which they are to be inserted and the length of the text file. The frequency of repetitive insertions of the same watermark should be such as to leave space for the insertion of the other types of watermarks.

One such long jumping rule is the following:
If N is the content of the counter when the last digit of the Watermark Body has been inserted then no further text characters should be watermarked so long as the condition

$$\text{(content of counter)} \leq N+F_r$$

holds, where $F_r$ is an integer depending on the frequency of repetitive insertion of the same watermark.

Thus, so long as long jumping rules are active the Text-Character Selection Procedure Unit 34 maintains the state of Line 35 at FALSE, which inhibits further watermarking of read text characters until such time as the content of the Counter 36 exceeds $N+F_r$. Once the counter content exceeds $N+F_r$ then the long jumping rules becomes inactive and the TCSP Unit 34 proceeds with the second insertion of the same specified watermark type based again on the short jumping rules which have returned to the active state. When the same specified watermark type has been

inserted throughout the text, a next specified watermark type is selected and, provided it follows the ordering rules, relating to the order in which watermark types are permitted to be inserted, the insertion process is repeated. Thus different watermark types are inserted in accordance with the position-division-multiplexing scheme.

In a preferred embodiment of the invention, if a selected text character is to be watermarked with a binary digit 0 which is part of the Body of the watermark to be embedded then the 'space' symbol is overwritten with itself X times and then overwritten with the selected text character. If the selected text character is to be watermarked with the digit 1 then the space symbol is overwritten with itself Y times and then overwritten with the selected character, where X≠Y.

It will be appreciated that the roles of binary zeroes and ones can be interchanged. The overwriting watermarking procedure, which leaves the actual text character unchanged, can be carried out using computer commands that are well known in the art.

Preferably, Portable Document Format (PDF) is used as the document format used when watermarking, as this can be used to represent a document in a manner independent of the application software, hardware and operating system used to create it.

Alternatively, a document formatted according to a particular word processing system may be watermarked by using software to change the basic commands of the word processing system to allow overwriting.

In another alternative method, the document could be converted from the format of a particular word processing system, such as Word (trade mark) or WordPerfect (trade

mark) into the PostScript language and then, if necessary into PDF using, for example, an Adobe Acrobat Distiller (registered trade marks).

To extract the watermark from the Output file, the file 14 must be input the Watermark Visualisation Unit 16. This first identifies the type of watermark(s) contained in the file and then, by performing generally the opposite procedure to that which was carried out by the Mapping Unit 23 and Encryption Unit 24 in order to encode the text with the watermark, the Visualisation Unit detects, reads and displays the watermark Body and Type contained within the text.

According to a second embodiment of the invention the binary zeros and ones representing the watermark are watermarked into the text by using X overwrites of a selected character by itself if the binary digit is a zero and by Y overwrites of the character by itself, (with X≠Y), if the binary digit is a one.

According to a third embodiment of the invention, binary zeros and ones are represented and watermarked into the text by respectively invisibly underlining or invisibly over-scoring the selected text character.

According to a fourth embodiment of the invention, the method is the same as in the first embodiment described except that a colourless character is used instead of the 'space' character.

According to a fifth embodiment of the invention, the method is the same as in the first embodiment described except that the 'null' character is used instead of the 'space' character.

According to a sixth embodiment of the invention, a binary zero to be embedded is represented by a colourless character, say P, overwritten by the selected text character, and a binary one is represented by a different colourless character, say Q, overwritten by the selected text character.

According to yet another embodiment of the invention pairs of binary digits taken from the binary sequence representing the watermark Body to be inserted are encoded as follows:

00 is encoded as a zero as in the first embodiment
01 is encoded as a one as in the first embodiment
10 is encoded as a zero as in the second embodiment
11 is encoded as a one as in the second embodiment

This process has the advantage of increasing the bandwidth efficiency of the watermarking procedure.

According to yet a further embodiment of the invention, binary zeros and ones, which are part of the body of the watermark to be inserted, or combinations of a number of binary digits, are encrypted and impressed on the formatted text file using combinations of characters that are invisible in the sense related to the previous embodiments.

The invention thus provides an improved method of watermarking which is insensitive to formatting changes, unlike known watermarking methods. Moreover, the watermarking is more secure and cannot be detected simply by scanning the watermarked text.

**Claims**

1. A method of inserting code data into a text data file representing a string of characters, comprising the steps of :

selecting a character to be coded with at least part of the code data,

reading at least part of the code data, and, in dependence on the code data read, replacing in the text data file the selected character by a plurality of characters, including the selected character, each of the plurality of characters being allocated the same position in the string of characters as the selected character.

2. A method according to claim 1, in which the steps of selecting a character, reading part of the code data, and replacing the selected character are repeated until all the code data has been inserted into the text data file.

3. A method according to claim 1 or 2, including the step of encrypting the code data to be inserted into the text data file.

4. A method according to claim 1, 2 or 3, including the step of storing the code data in one or more buffers, and reading at least part of the code data from a buffer.

5. A method according to claim 4 in which the code data is stored as a string of symbols, successive symbols being read from the buffer as successive selected characters are coded.

6. A method according to claim 5 in which the symbols are read in pairs or groups.

7. A method according to any preceding claim, in which said plurality of characters includes the 'space' or 'null' character or a combination thereof.

8. A method according to any preceding claim, in which said plurality of characters includes colourless characters.

9. A method according to claim 8, in which said plurality of characters includes an underlining or overscoring character.

10. An apparatus for inserting code data into a text data file representing a string of characters, comprising:

means for selecting a character to be coded with at least part of the code data; and

means for replacing the selected character in the text data file by a plurality of characters, including the selected character, with each of the plurality of characters being placed at the same position in the string of characters as the selected character, said plurality of characters being chosen in dependence on the code data read.

11. An apparatus according to claim 10, including means for encrypting the code data to be inserted into the text data file.

12. An apparatus according to claim 10 or 11, including one or more buffers for storing the code data.

13. An apparatus according to claim 10, 11 or 12, including means for identifying whether a text data file has code data inserted therein.

14. A method of inserting code data into a text data file substantially as herein described.

15. An apparatus for inserting code data into a text data file substantially as herein described with reference to the drawings.

The Patent Office

15

| | | | |
|---|---|---|---|
| **Application No:** | GB 9815300.0 | **Examiner:** | B.J.SPEAR |
| **Claims searched:** | 1-15 | **Date of search:** | 8 February 1999 |

## Patents Act 1977
## Search Report under Section 17

**Databases searched:**

UK Patent Office collections, including GB, EP, WO & US patent specifications, in:

UK Cl (Ed.Q): H4F (FBB); H4P (PDCSA, PDCSK, PDCSX)

Int Cl (Ed.6): G06F 1/00; H04N 1/32

Other: Online:WPI,INSPEC

**Documents considered to be relevant:**

| Category | Identity of document and relevant passage | | Relevant to claims |
|---|---|---|---|
| A | EP0798619A2 | (Sun Microsystems) | - |
| A | US5467447 | (Vogel) | - |

| | | | |
|---|---|---|---|
| X | Document indicating lack of novelty or inventive step | A | Document indicating technological background and/or state of the art. |
| Y | Document indicating lack of inventive step if combined with one or more other documents of same category. | P | Document published on or after the declared priority date but before the filing date of this invention. |
| & | Member of the same patent family | E | Patent document published on or after, but with priority date earlier than, the filing date of this application. |